Developing and Validating Measurement Scales During Pandemic Conditions: A Case Study with the Scale for Habitat Usability

Ian Robertson[1,3], Brandin Munson[2,3], Jerri Stephenson[3], Ryan Z. Amick[1,3]
[1]KBR, Houston, TX
[2] University of Houston, Houston, TX
[3] NASA Johnson Space Center, Houston, TX

At NASA, habitat evaluations often employ subjective measures. Some measures are frequently used, well-established tools, whereas others are homegrown measures tailored to specific projects. The variety of measures used makes evaluation comparisons across projects difficult. Additionally, some of these measures are burdensome, may be too specialized, or may require an expert to use and interpret, limiting their utility. Taken together, these drawbacks suggest the need for a new measurement tool. To that purpose, a team at NASA worked on developing a new scale for measuring habitat usability, the Scale for Habitat Usability (SHU). The SHU is intended to be a quick, multi-faceted measure for evaluating habitat usability across the development lifecycle. However, like many research projects, the development of the SHU faced setbacks due to the COVID-19 pandemic. Pandemic prevention protocols precluded in-person data collection, forcing the team to take *some* non-traditional approaches to scale development. This paper reports the steps the team took to complete the project.

Introduction

When designing space habitats and vehicles, it is imperative to incorporate data from representative users. Often, users' input is captured with subjective measurement scales. Using this data, iterative design changes may be made which accurately address astronaut crew needs and pain points. However, it is difficult to do this in a consistent and comparable way across many teams. Many of the evaluation tools currently employed by NASA teams can be time-consuming to administer, are inflexible to specific aspects of the habitat, and many require experts to effectively collect data. Therefore, the goal of this multi-phase study is to create and validate a habitat usability survey tool which is flexible and easily administered across a variety of mockup types.

Prior Work

The first phase of this study consisted of the creation of a preliminary habitat usability scale, the Scale for Habitat Usability (SHU). The SHU was created with input from habitat designers and human factors engineering subject matter experts (SMEs), and incorporated items from habitat assessment and usability questionnaires. The preliminary survey was then refined following recommended methods as identified by Hinkin (2005), resulting in a survey with nine subscale constructs (Cognitive Workload, Labels/Wayfinding, Layout, Lighting, Likeability, Intuitiveness, Physical Workload, Situation Awareness, Usability), with a total of 54 survey items. Note that the survey was created for the assessment of a habitat's usability in relation to the performance of a task (e.g., the user rates the habitat/vehicle as it relates to a task of interest).

Current Work

The purpose of the second phase of the study was to refine the item pool and to collect evidence regarding the reliability and validity of the SHU. This was completed by collecting expert feedback through 1) an open card sort, 2) a closed card sort, and 3) conducting a prospective usability evaluation to collect data to assess the scale's reliability and validity. These steps further refined the survey to result in a more valid tool and serve as necessary steps towards establishing the final instrument.

Refinement of the Item Pool

In scale development, it is considered best practice to incorporate expert feedback when determining how well questionnaire items represent the domain of interest (Boateng et al., 2018; Carpenter, 2018). To this end, card sorting was used to analyze the content validity of the item pool (Capra, 2005). Both an open sort and a closed sort were conducted. The open sort was completed first, and its results were used to refine the item pool prior to the closed sort. The rationale behind the card sorting is that items that belong to the same construct should be reliably sorted into the same content group. The open card sort leveraged experts to identify what constructs are measured by the candidate items. That is, participants were given survey items with no predetermined construct groupings and were asked to sort them into groups of their own making. The closed card sort tested the reliability of the constructs identified from the open card sort. That is, participants in the closed sort were restricted to sort items into construct groups identified in the open sort.

Open Card Sort

Participants

Twenty-four participants were recruited to complete the card sort. Participants' responses to the item "Please list relevant area(s) of expertise you possess" were screened to ensure they met inclusion criteria set for the study (expertise pertaining to architecture, habitat design, or human factors). After screening, 21 participants were retained for the analysis. The mean reported age was 49.75 years old (*SD* = 11.37).

Participants reported their gender as female (*N* = 12) and male (*N* = 9). Participants' self-reported level of education included doctorate degree (*N* = 13), master's degree (*N* = 6), and bachelor's degree (*N* = 2).

Materials

The open card sort was conducted with a pool of 71 items (additional items had been added since the first phase) intended to measure various aspects of the usability of built environments. The open sort was hosted on the website www.provenbyusers.com.

Procedure

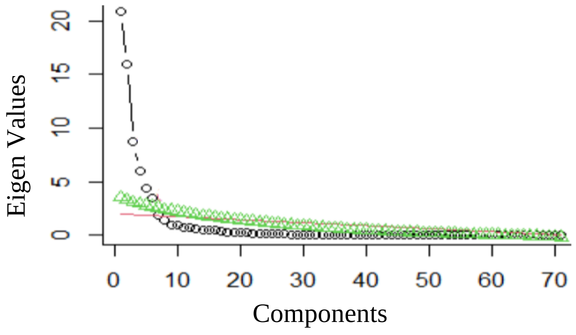
Eligible potential participants were sent an email explaining the goals of the study. Potential participants who expressed interest in taking part of the study were sent a link to the study. Participants were instructed to read each item and group it into categories of their own creation. They repeated this process until all items were exhausted.

Results

On average, participants sorted the cards into 8 groups as measured by median (*SD* = 2.2). Participants took approximately one hour and nineteen minutes to complete the sort. This time appears to be biased by extreme values (more than 7 hours) which may be due to participants leaving the card sort open in a browser window. The average time to complete the sort was 41.25 minutes (*SD* = 25.66) when these extreme observations were removed. A parallel analysis suggested that a six-factor solution fit the data best (see Figure 1).

Figure 1

Scree Plot and Parallel Analysis for the Open Card Sort



We performed a principal axis factor analysis with an oblique rotation on the similarity matrix from the participants' card sorts. We inspected the six-, seven-, and eight-factor solutions. The factor loadings across each solution were of similar strength. We selected the seven-factor solution because we judged that it replicated the highest number of the original nine subscale constructs created in Phase 1 while still maintaining cogency among the items belonging to each factor. That is, items within factors were more similar in content to one another and were consistent with factor definitions. The seven-factor solution explained 58% of the observed variance (six-factor = 54%, eight-factor = 60%). A summary of the factor solution can be found in Table 1.

Table 1

Open Sort Summary Statistics for the Seven-Factor Solution

Factor	# of Items	*M* Factor Loading	Variance Explained
Intuitiveness	10	.60	7%
Labels	5	.75	7%
Layout	15	.63	12%
Lighting	9	.95	10%
Satisfaction	11	.55	7%
Situation awareness	9	.59	5%
Workload	12	.67	10%

Outcomes

Based on the results of the seven-factor solution, five items were dropped from the item pool due to cross-loading on multiple factors. One item ("I am satisfied with the support information for using this habitat.") was retained but was reworded ("Information available throughout the habitat was sufficient to complete the task.") to better fit the construct group it was sorted into (Labels).

Closed Card Sort

Participants

In total, 7 participants were recruited to complete the closed card sort. The mean reported age was approximately 43.43 years old (*SD* = 11.39). Participants self-identified gender as female (*N* = 4) and male (*N* = 3). Participants' self-reported level of education included doctorate degree (*N* = 3), bachelor's degree (*N* = 3), and master's degree (*N* = 1).

Materials

The closed sort used the items retained from the open sort. Like the open sort, the closed sort was hosted on the website www.provenbyusers.com.

Procedure

Eligible potential participants were sent an email explaining the goals of the study. Potential participants who expressed interest in taking part of the study were sent a link to www.provenbyusers.com. Participants were instructed to read each item and group it into what they judged to be the most appropriate category. Unlike the open sort, the group labels (the same categories listed in Table 1) were provided in the closed sort. They repeated this process until all items were exhausted.

Results

On average, participants took 13.16 minutes (*SD* = 5.78) to complete the closed card sort. Because the closed card sort had a predetermined number of content groups, we only inspected the seven-factor solution. We performed a principal axis factor analysis with an oblique rotation on the similarity matrix from the participants card sorts. The seven-factor solution explained 72% of the observed variance. A summary of the factor solution statistics can be found in Table 2.

Table 2

Summary Statistics for the Closed Sort Seven-Factor Solution

Factor	# of Items	*M* Factor Loading	Variance Explained
Intuitiveness	10	.77	11%
Labels	4	.93	6%
Layout	16	.80	17%
Lighting	9	.94	12%
Satisfaction	11	.73	11%
Situation awareness	5	.75	5%
Workload	11	.69	9%

Outcomes

Five items were cut from the item pool due to cross-loading or poor replication of their performance in the open sort. In total, 61 items were retained for the next stage.

Prospective Usability Evaluation

To further develop the SHU, it was necessary to conduct a study that could approximate a physical habitat evaluation with representative users while providing enough data to perform item refinement. However, due to COVID-19 restrictions on in-person studies, we used non-traditional methods to approximate habitat evaluations. *Prospective inspection*, where the usability of a system is rated before use, has been suggested as an alternative evaluation method when in-person testing in not an option (Robertson & Kortum, 2020). This method has the benefit of approximating a low fidelity evaluation, a use case that is relevant to the SHU, as users often do not interact directly with low-fidelity mockups. Additionally, there is evidence that usability ratings of built environments can be collected by combining visual media (photos and videos) with task descriptions. This method, often referred to as photo elicitation, has been used in several domains (e.g., sociology, environmental psychology, landscaping) to collect data about users' preferences for different environments (Alexander, 2013; Stamps, 1990; Xiang et al., 2021).

There is additional evidence to support photo elicitation for use in evaluating the usability of environments. In one study, photorealistic renderings were used to assess how environmental variables (e.g., spatial configuration, presence of dividers) impact the perceived usability of voting machines using the System Usability Scale (SUS; Acemyan & Kortum, 2016). Acemyan and Kortum (2018) found practically insignificant (as measured by effect size) or no differences in users' ratings of usability and acceptability of two built environments, a polling station and classroom, between presentation mediums (in person, photorealistic renderings and photographs).

The major drawback to this approach is that prospective usability ratings tend to be higher on average than ratings based on actual use (Robertson & Kortum, 2019; Robertson & Kortum, 2020). However, this drawback has little bearing on this study. The outcomes of interest are not the actual ratings but the inter-relationships between items on the SHU (e.g., scale dimensionality and reliability), as well as their relationship to relevant usability outcomes (construct validity).

Additionally, it has been suggested that prior experience with a product (e.g., expertise) may lessen this drawback (Robertson & Kortum, 2019; Robertson & Kortum, 2020). Visual media have also been found to improve the quality of inspection methods. In one study, experts who conducted cognitive walkthroughs supplemented with video found as many problems as experts who conducted a cognitive walkthrough in situ. Both the video and in situ groups found more problems than experts who conducted a traditional cognitive walkthrough (Gabrielli et al., 2005). This suggests that the drawbacks of prospective usability evaluation *may* be mitigated using experienced participants and augmenting the evaluation with visual stimuli.

Study Goals

The goals of the study were to further reduce the total number of items, estimate the reliability of the SHU subscales and to provide initial validity data for the SHU.

Method

Participants

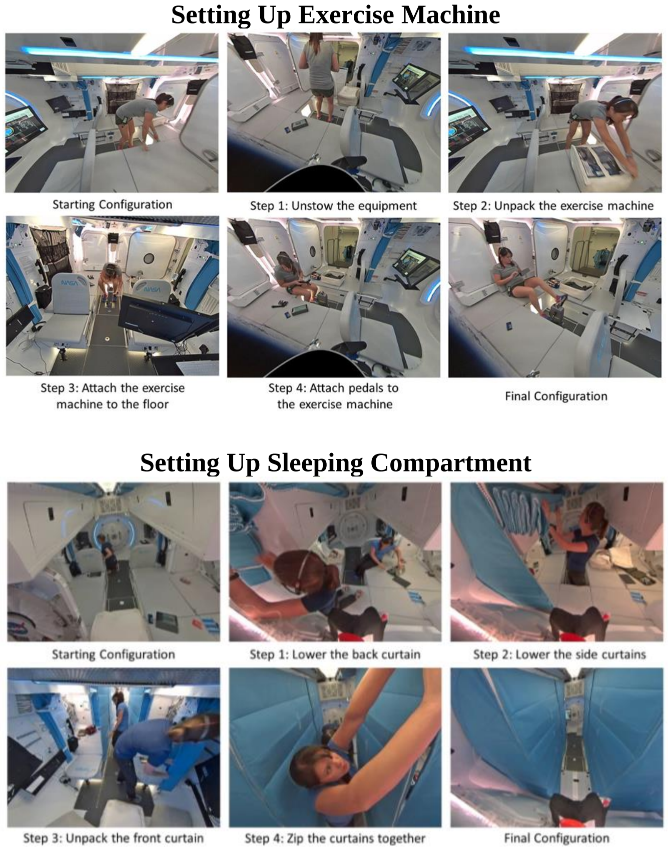
Thirty-four participants from NASA's Human Test Subject Facility pool were recruited for the study. Participants' average reported age was 39.75 years old (*SD* = 7.52). Participants' reported gender included male (*N* = 19) and female (*N* = 15). In terms of education, participants reported having earned a master's degree (*N* = 20), a bachelor's degree (*N* = 6), some post undergraduate work (*N* = 3), a doctorate degree (*N* = 3), a specialist's degree (*N* = 1), and an applied or professional doctorate degree (*N* = 1).

Design and Materials

The study used a repeated measures design. In the study, participants prospectively evaluated an environment for two different tasks. Participants rated the environment for setting up an exercise machine or a sleeping compartment (see Figure 3) based on photographs of the environment and a representative user completing the task.

Figure 2

Task Procedures and Images for the Two Tasks


Setting Up Exercise Machine
Starting Configuration
Step 1: Unstow the equipment
Step 2: Unpack the exercise machine
Step 3: Attach the exercise machine to the floor
Step 4: Attach pedals to the exercise machine
Final Configuration
Setting Up Sleeping Compartment
Starting Configuration
Step 1: Lower the back curtain
Step 2: Lower the side curtains
Step 3: Unpack the front curtain
Step 4: Zip the curtains together
Final Configuration

The presentation order of the tasks was counterbalanced so that half the participants rated the environment for the sleeping compartment first and the exercise machine second and vice versa. The presentation order was randomly assigned.

Participants rated the environment using the items retained from the closed card sort. For purposes of validation, participants were asked to rate the habitat in terms of overall usability using the adjective ratings scale (Bangor et al., 2009). To measure perceived efficiency, participants were asked to estimate the amount of time they think it would take to complete the task in minutes: "About how long do you think it would take you to complete this task?". To measure perceived effectiveness, participants were asked to rate the likelihood of their success in completing the task: "I think I would be able to complete the task successfully.". Both questions were rated on a Likert scale of 1 (Strongly Disagree) to 5 (Strongly Agree). The study was hosted on www.provenbyusers.com.

Results

Item Reduction

Items were selected using alpha-if-item deleted, inter-item correlations, adjusted item-total correlations, and practical/theoretical relevance as judged by the scale developers (Boateng et al., 2018). Practical relevance was judged by how applicable an item may be across the XR spectrum. In total, 28 items were retained, and one was reworded. After reducing the number of items, the subscales still met the recommended threshold for internal reliability. Table 3 shows the scale scores and reliability.

Table 3

Subscale Scores and Reliability by Task

Subscale	*M*	*SD*	α [95% CI]
	Exercise Machine		
Intuitiveness	4.06	0.69	.79 [.68, .91]
Labels	3.38	0.96	.96 [.94, .98]
Layout	3.92	0.7	.83 [.74, .92]
Lighting	4.13	0.76	.92[.88, . 97]
Satisfaction	4.05	0.57	.82 [.72, .92]
Situation Awareness	3.95	0.67	.81 [.70, .91]
Workload	4.41	0.61	.89 [.82, .96]
Total average	3.99	0.55	–
	Sleeping Quarters		
Intuitiveness	4.39	0.55	.80 [.70, .90]
Labels	3.46	1.01	.95 [.93, .98]
Layout	3.98	0.71	.85 [.77, .93]
Lighting	4.3	0.65	.88 [.82, .95]
Satisfaction	4.23	0.62	.85 [.77, .93]
Situation Awareness	4.15	0.61	.84 [.75, .92]
Workload	4.51	0.56	.93 [.89, .97]
Total average	4.14	0.53	–

Notes. *N* = 34.

Validity

As seen in Table 4, most participants rated the usability as of the habit for both tasks as "OK" or better using the adjective rating scale.

Table 4

Adjective Usability Rating by Task

Adjective	Exercise Machine (*N* = 33)	Sleeping Quarters (*N* = 34)
Best Imaginable	3%	0%
Excellent	21%	38%
Good	48%	56%
OK	24%	6%
Poor	3%	0%
Awful	0%	0%
Worst Imaginable	0%	0%

Note. Responses to "Overall, I would rate the user-friendliness of this product as . . .".

To provide initial evidence for the validity of the SHU, the total SHU scores (computed by averaging the subscale scores) were correlated with the adjective rating scale, estimated task success, and estimated time on task. As seen in Table 5, the validation coefficients for the adjective rating scale and estimated success were statistically significant and adequately strong for purposes of construct validity (Nunnally, 1978, p. 90) and made theoretical sense (high success is associated with high usability). For both tasks the validity coefficient for time was weak and not statistically significant.

Table 5

Validity Coefficients for both Tasks

Criterion	*r* [95% CI]	Magnitude
	Exercise Machine	
Adjective Rating	.42 [.09, .67]*	Medium/Large
Success	.40 [.07, .66]*	Medium/Large
Time	-.15 [-.48, .23]	Small
	Sleeping Quarters	
Adjective Rating	.46 [.15, .69]**	Medium/Large
Success	.49 [.18, .71]**	Medium/Large
Time	-.08 [-.42, .28]	Small

Notes. * $p < .05$, ** $p < .01$.

DISCUSSION

The goal of this project was to create a flexible, reliable, and valid measure of habitat usability. Given the circumstances created by COVID-19, the scale developers were forced to make some nontraditional choices in their approach to the scale development process. For example, in lieu of factor analysis, scale dimensionality was largely determined via card sorts conducted with experts. Because in-person testing was restricted, the developers were forced to collect responses online. While this approach was not preferred, it was partially supported by the literature. To ensure that the tool and its subscale components are valid and reliable across mockup types, it is important to continue the iterative refinement process. Future work will seek to validate the subscales and to collect additional evidence for the reliability of the subscales with in-person studies using physical and VR mockups.

REFERENCES

Acemyan, C. Z., & Kortum, P. (2016). Does the polling station environment matter? The relation between voting machine layouts within polling stations and anticipated system usability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 1066-1070. https://doi.org/10.1177/1541931215591299

Acemyan, C. Z., & Kortum, P. (2018). Does the type of presentation medium impact assessments of the built environment? An examination of environmental usability ratings across three modes of presentation. *Journal of Environmental Psychology*, *56*, 30-35. https://doi.org/10.1016/j.jenvp.2018.02.006

Alexander, V. D. (2013). Views of the neighbourhood: A photo-elicitation study of the built environment. *Sociological Research Online*, *18*(1), 1-26. https://doi.org/10.5153/sro.2832

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114-123. https://doi.org/10.5555/2835587.2835589

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Meglar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontier in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P.W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189-194). Taylor & Francis. https://doi.org/10.1201/9781498710411-35

Capra, M. G. (2005). Factor analysis of card sort data: An alternative to hierarchical cluster analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(1), 691-695. https://doi.org/10.1177/154193120504900512

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers, *Communication Methods and Measures*, *12*(1), 25-44. https://doi.org/10.1080/19312458.2017.1396583

Gabrielli, S., Mirabella, V., Kimani, S., & Catarci, T. (2005). Supporting cognitive walkthrough with video data: A mobile learning evaluation study. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services* (pp. 77-82). https://doi.org/10.1145/1085777.1085791

Hinkin, T. R. (2005). Scale development principles and practices. In R. A. Swanson & E. F. Holton (Eds)., *Research in organizations: Foundations and methods of inquiry* (pp. 161-179). Berrett-Koehler. https://www.drrichardgreen.com/uploads/3/4/5/2/34520924/swanson_and_holton-2005-research_in_organizations.pdf#page=180

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Robertson, I., & Kortum, P. (2019). An investigation of different methodologies for rating product satisfaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1259-1263. https://doi.org/10.1177%2F1071181319631071

Robertson, I., & Kortum, P. (2020). Validity of three discount methods for measuring perceived usability. *Journal of Usability Studies*, *16*(1), 13-28. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Robertson_Nov2020.pdf

Stamps, A. E. (1990). Use of photographs to simulate environments: A meta-analysis. *Perceptual and Motor Skills*, *71*, 907-913. https://doi.org/10.2466/PMS.71.7.907-913

Xiang, Y. Liang, H., Fang, X., Chen, Y., Xu, N., Hu, M., Chen, Q., Mu, S., Hedblom, M., Qiu, L., & Gao, T. (2021). The comparisons of on-site and off-site applications in surveys on perception of and preference for urban green spaces: Which approach is more reliable? *Urban Forestry and Urban Green*, *58*, 126961. https://doi.org/10.1016/j.ufug.2020.126961